Sociology 63993
Exam 2 Answer Key - DRAFT
March 27, 2009

I. True-False. (20 points) Indicate whether the following statements are true or false. If false, briefly explain why.

1. A researcher regresses Political Liberalism (a scale that ranges between 0 and 100) on X. She does not include dummy variables or interaction terms for race. If the model is correct, it means that, on average, blacks and whites are equally liberal.

False. Unless blacks and whites have the same mean value for X, they will have different mean values for Political Liberalism.

2. A researcher has inadvertently omitted an important variable from her model, resulting in omitted variable bias. Unfortunately, increasing the sample size will not help to reduce this bias.

True.

3. A researcher runs the following regressions:

```
. reg health weight if white

      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  1, 10333) =   18.08
       Model |  26.2659433     1  26.2659433           Prob > F      =  0.0000
    Residual |  15008.7554 10333  1.45250706           R-squared     =  0.0017
-------------+------------------------------           Adj R-squared =  0.0017
       Total |  15035.0214 10334   1.4549082           Root MSE      =  1.2052

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0032831   .0007721    -4.25   0.000    -.0047965   -.0017698
       _cons |   3.649905   .0567655    64.30   0.000     3.538634    3.761176
------------------------------------------------------------------------------

. reg health weight if !white

      Source |       SS       df       MS              Number of obs =     516
-------------+------------------------------           F(  1,   514) =    2.05
       Model |  2.96522291     1  2.96522291           Prob > F      =  0.1523
    Residual |  741.707258   514  1.44301023           R-squared     =  0.0040
-------------+------------------------------           Adj R-squared =  0.0020
       Total |  744.672481   515  1.44596598           Root MSE      =  1.2013

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0049025     .00342    -1.43   0.152    -.0116213    .0018164
       _cons |   3.707847   .2493572    14.87   0.000     3.217962    4.197731
------------------------------------------------------------------------------
```

As the results show, weight has a statistically significant effect on the health of whites, but the effect is not significant for nonwhites. The researcher should therefore conclude that the effect of weight on health is significantly greater for whites.

False. The differences in statistical significance may just reflect the fact that the white sample is more than 20 times as large as the nonwhite sample. You would need to do a formal test to see whether the effects significantly differ. Further, the estimated effect

of weight is actually larger for nonwhites than it is for whites, which is the opposite of what the statement claims.

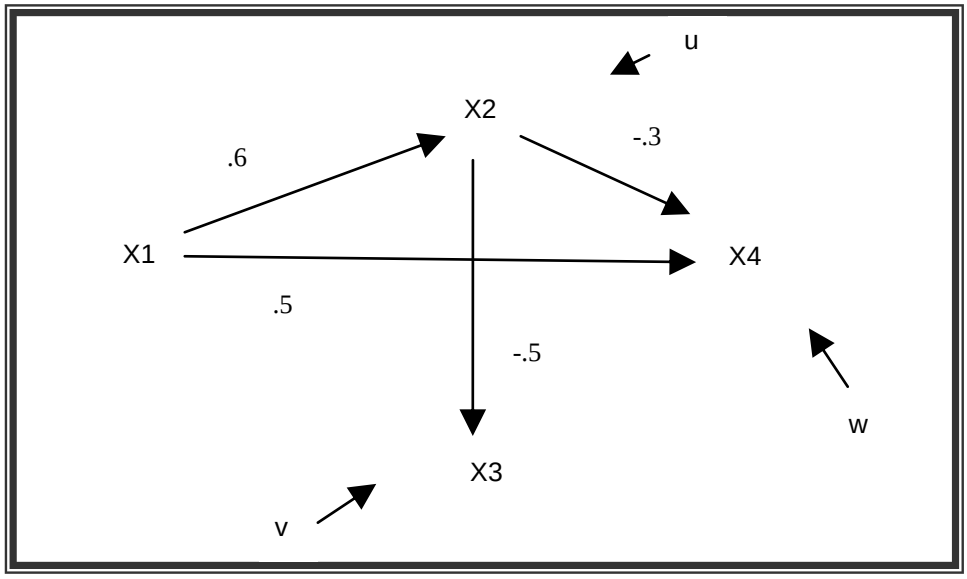
4. Life satisfaction (measured on a 50 point scale) is regressed on Income, Female, and Female*Income. The coefficient for Female is +5. This means that, whenever a man and a woman have equal incomes, the woman is expected to score 5 points higher than the man on life satisfaction.

False. Because the model includes an interaction term, 5 points is the expected difference only for a man and woman who both have 0 income. For other values of income, the expected difference will be smaller or greater than 5. The statement would be true if the model did not include the interaction term.

5. Exponential models are appropriate when we believe that the relationship between two variables is curvilinear.

False. Polynomial models are called for in such cases.

II. Path Analysis/Model specification (25 pts). A sociologist believes that the following model describes the relationship between X1, X2, X3, and X4. All her variables are in standardized form. The estimated value of each path in her model is included in the diagram.



a. (5 pts) Write out the structural equation for each endogenous variable, using both the names for the paths (e.g. β_{42}) and the estimated value of the path coefficient.

$X2 = \beta_{21}X1 + u = .6 * X1 + v$

$X3 = \beta_{32}X2 + u = -.5 * X2 + v$

$X4 = \beta_{41}X1 + \beta_{42}X2 + w = .5 * X1 - .3* X2 + w$

b. (10 pts) Part of the correlation matrix is shown below. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

```
             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |    .6000     1.0000
          x3 |      ?         ?     1.0000
          x4 |      ?         ?        ?     1.0000
```

Using normal equations,

$r12 = \beta_{21}E(X1^2) = .6$

$r13 = \beta_{32}E(X1X2) = -.5*.6 = -.3$

$r23 = \beta_{32}E(X2^2) = -.5$

$r41 = \beta_{41}E(X1^2) + \beta_{42}E(X1X2) = .5 + (-.3 * .6) = .32$

$r42 = \beta_{41}E(X1X2) + \beta_{42}E(X2^2) = (.5 * .6)\ \ -.3 = 0$

$r43 = \beta_{41}E(X1X3) + \beta_{42}E(X2X3) = (.5 * -.3) + (-.3 * -.5) = 0$

	X1	X2	X3	X4
X1	1.00			
X2	0.60	1.00		
X3	-.3	-.50	1.00	
X4	0.32	0.00	0.00	1.00

c. (5 pts) Decompose the correlation between X1 and X4 into

- Correlation due to direct effects

  .5

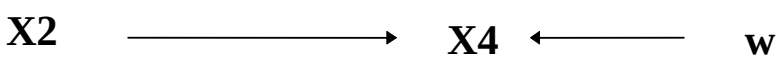
- Correlation due to indirect effects

  -.18

- Correlation due to common causes

  0

d. (5 pts) Suppose the above model is correct, but instead the researcher believed in and estimated the following model:



What conclusions would the researcher likely draw? In particular, what would the researcher conclude about the effect of changes in X2 on X4? Discuss the consequences of this mis-specification, and in what ways, if any, the results would be misleading. Why would she make these mistakes?

The estimated effect would be the same as the correlation between the two variables, i.e. 0. The researcher would therefore conclude that changes in X2 would have no effect on X4, when in reality increases in X2 result in decreases in X4. This mistake arises from the fact that correlation that is due to the common cause of X1 is instead treated as correlation due to direct effect.

III. Group comparisons (25 points). It is mid-April 2009. To the dismay of Notre Dame officials, the controversy over having Barack Obama as commencement speaker continues to rage. More than 500,000 people have signed an online petition protesting the invitation. Hundreds of alumnae have withdrawn their pledges to the University, while dozens of parents are threatening to boycott their own child's graduation ceremony. At the same time, thousands of alumnae and students have expressed strong support for the decision. With the University's finances already suffering, administrators desperately feel that they need to better understand who is supporting the University, and why. An outside polling firm has therefore collected information from more than 10,000 ND alumnae on the following variables:

Variable	Description
nd	Likelihood of donating to Notre Dame, measured on a scale that runs from -100 to +100
prolife	Importance of prolife issues to the respondent. The original item was measured on a scale that ranges from 0 to 200, but the measure used in the analysis has been centered to have a mean of zero.
dem	Coded 1 if the respondent is a Democrat, 0 if Republican
demlife	dem * prolife

The results of the analysis are as follows:

```
. * See if there are differences in support by party affiliation
. ttest nd, by(dem)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4909    13.47488    .3632543    25.45114    12.76274    14.18702
       1 |    5426    49.66043    .3191398    23.50828    49.03479    50.28607
---------+--------------------------------------------------------------------
combined |   10335    32.47273    .2990589    30.40269    31.88652    33.05895
---------+--------------------------------------------------------------------
    diff |           -36.18555    .4816196               -37.12961   -35.24148
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -75.1330
Ho: diff = 0                                     degrees of freedom =    10333

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000


. * Estimate Models
. nestreg:  reg nd prolife dem demlife
```

Block 1: prolife

```
      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  1, 10333) = 1486.87
       Model |  1201578.31     1  1201578.31           Prob > F      =  0.0000
    Residual |  8350378.41 10333  808.127205           R-squared     =  0.1258
-------------+------------------------------           Adj R-squared =  0.1257
       Total |  9551956.71 10334  924.323274           Root MSE      =  28.428

------------------------------------------------------------------------------
          nd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prolife |  -.7022148    .018211   -38.56   0.000    -.7379119   -.6665177
       _cons |   32.47431   .2796306   116.13   0.000     31.92618    33.02244
------------------------------------------------------------------------------
```

Block 2: dem

```
      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  2, 10332) = 3075.43
       Model |  3564476.12     2  1782238.06           Prob > F      =  0.0000
    Residual |  5987480.59 10332  579.508381           R-squared     =  0.3732
-------------+------------------------------           Adj R-squared =  0.3730
       Total |  9551956.71 10334  924.323274           Root MSE      =  24.073

------------------------------------------------------------------------------
          nd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prolife |  -.3013264   .0166504   -18.10   0.000    -.3339643   -.2686885
         dem |   32.69198   .5119748    63.85   0.000     31.68841    33.69555
       _cons |   15.30973   .3582313    42.74   0.000     14.60752    16.01193
------------------------------------------------------------------------------
```

Block 3: demlife

```
      Source |       SS       df       MS              Number of obs =   10335
-------------+------------------------------           F(  3, 10331) = 2092.17
       Model |  3609987.84     3  1203329.28           Prob > F      =  0.0000
    Residual |  5941968.87 10331   575.15912           R-squared     =  0.3779
-------------+------------------------------           Adj R-squared =  0.3778
       Total |  9551956.71 10334  924.323274           Root MSE      =  23.982

------------------------------------------------------------------------------
          nd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     prolife |  -.4689101   .0251012   -18.68   0.000    -.5181134   -.4197068
         dem |   32.38671   .5112032    63.35   0.000     31.38465    33.38876
     demlife |   .2975046   .0334446     8.90   0.000     .2319467    .3630626
       _cons |   16.33018   .3748686    43.56   0.000     15.59537      17.065
------------------------------------------------------------------------------
```

Block	F	Block df	Residual df	Pr > F	R2	Change in R2
1	1486.87	1	10333	0.0000	0.1258	
2	4077.42	1	10332	0.0000	0.3732	0.2474
3	79.13	1	10331	0.0000	0.3779	0.0048

```
. * Finally, test for differences in prolife attitudes by party affiliation
. ttest prolife, by(dem)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4909    6.089225     .194648    13.63787    5.707628    6.470822
       1 |    5428   -5.506997    .1999352    14.73022    -5.89895   -5.115044
---------+--------------------------------------------------------------------
combined |   10337    2.37e-06    .1510277    15.35515   -.2960412    .2960459
---------+--------------------------------------------------------------------
    diff |            11.59622    .2801171                11.04714    12.14531
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  41.3978
Ho: diff = 0                     degrees of freedom =    10335

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

The initial t-test shows that Democrats are significantly more likely to donate to Notre Dame. Based on the remaining results, explain to the Notre Dame administration why that is the case. When thinking about your answers, keep in mind the various reasons that two groups can differ on some outcome measure. Specifically, answer the following:

a) (15 pts) The researchers estimate a series of models. Which of the models do you think is best, and why? What do these models tell us about how concern about prolife issues affects the likelihood of donating to the University? What ways (if any) do the determinants of support for Notre Dame differ by party affiliation? What insights, if any, does this give us as to why Democrats tend to be more supportive of Notre Dame?

The final model, which allows both the intercepts and slopes to differ by party affiliation, is best (i.e. all terms are statistically significant). The model shows us that those who are more strongly prolife are less likely to support the university; however, the effect is significantly smaller for Democrats than it is for Republicans. This suggests that Democrats are more supportive of the University because their prolife attitudes have less of an impact on their support. Further, because prolife is centered, the coefficient for Democrat shows us that the "average" Democrat scores more than 32 points higher in support for the University than does the "average" Republican.

b) (10 pts) The researchers then do one last t-test. What does this test tell us about how the pro-life attitudes of alumnae differ by party affiliation? What additional insights, if any, does this test give us as to why Democrats are more supportive of Notre Dame?

The results show that Democrats are less strongly prolife than are Republicans. Thus, not only do prolife attitudes have less of an impact on Democrats, the fact that they have lower average scores on prolife further contributes to their higher levels of support.

IV. Short answer. Answer *both* of the following questions. (15 points each, 30 points total.) Each of the following describes a nonlinear or nonadditive relationship between variables. Draw a scatterplot that illustrates the relationship. Describe the harms that might result if you simply regressed Y on X, e.g. would values be over-estimated, under-estimated, or what? Indicate the model you think should be estimated, e.g. $E(Y) = \alpha + \beta_1 X + \beta_2 X^2$. Explain what variables you would need to compute in order to actually estimate the model, e.g. logs of variables, interaction terms. Finally, indicate how you would actually test whether or not nonlinearity or nonadditivity actually was a problem. If you find it helpful, you are welcome to present the Stata commands you would use, but the statistical rationale behind the command still needs to be clear.

a. The Director of Graduate Studies is concerned about advanced students dropping out of the program; and if they are going to drop out, he wonders why they do not do so sooner. He theorizes that student satisfaction steadily increases during the first four years of study, as students take classes and prepare for area exams. However, after that, as students work on their dissertation, their level of satisfaction steadily decreases across time.

This suggests a curvilinear relationship. A polynomial model that regresses both year and year^2 could be estimated. If the effect of year^2 is not significant, the theory would be rejected. (And of course the effects would have to be in the predicted direction). Stata's ovtest command could also be used after regressing satisfaction on year.

Another possibility would be a spline function. Year would be allowed to have one effect for years 1-4, and a different effect after that. The first slope would need to be positive while the second was negative. You could test this model by testing whether or not the two slopes were the same.

With either of the above, you would use Wald tests, although you could also set them up as incremental F test problems.

If you just ran a straight linear regression, you might find that there was little or no effect of year of study on satisfaction; the positive effects at the early stages would be offset by the negative effects at the later stages.

b. Conservative faculty at Georgetown feel that the Obama controversy at Notre Dame gives Georgetown a golden opportunity to finally stake its rightful claim to being the premier Catholic University in America. To add to Notre Dame's woes, it is sending its own pro-life literature to ND alumnae. Because of imperfect distribution methods, some ND alumnae will no doubt receive and read more of this literature than will others. The Georgetown professors feel that exposure to their literature will make alumnae more critical of Notre Dame, but they also think the effect will be greater for men than it is for women.

The model calls for an interaction effect, i.e. you would regress attitudes toward ND (let's assume higher values indicate greater support for the University) on exposure to pro-life literature, gender, and gender * exposure. If the Georgetown conspirators are correct, the main effect of exposure will be negative while the interaction term will be positive but still smaller than the main effect. You would estimate a sequence of models similar to what was done in Part III in order to determine what differences, if any, existed in the models for men and women. In any event, we can rest assured that their dastardly plan can never be successful in achieving its goal of knocking Notre Dame out of first place.

Again, Wald tests or incremental F tests could be used.

If you ignored the interaction terms, the effect of exposure would likely be under-estimated for men and over-estimated for women.

[NOTE: Answers should also include scatter plots depicting the relationships.]

APPENDIX: Stata Code for Selected Problems

```
* Part II – Path Analysis
clear
matrix input corr = (1,.6,-.3,.32\.6,1,-.5,0\-.3,-.5,1,0\.32,0,0,1)
corr2data x1 x2 x3 x4, corr(corr) n(100)
corr x1 x2 x3 x4
reg x2 x1
reg x3 x2 x1
reg x4 x1 x2 x3



* Part III – Interaction effects
* Generate the variables by manipulating nhanes2f
webuse nhanes2f, clear
keep health weight female
center weight, gen(prolife)
clonevar dem = female
gen nd = health * -20 - .4*prolife + 30 * dem + 85
gen demlife = dem * prolife
* See if there are differences in support by party affiliation
ttest nd, by(dem)
* Estimate Models
nestreg:  reg nd prolife dem demlife
* Finally, test for differences in prolife attitudes by party affiliation
ttest prolife, by(dem)
```